Assignment-based Subjective Questions

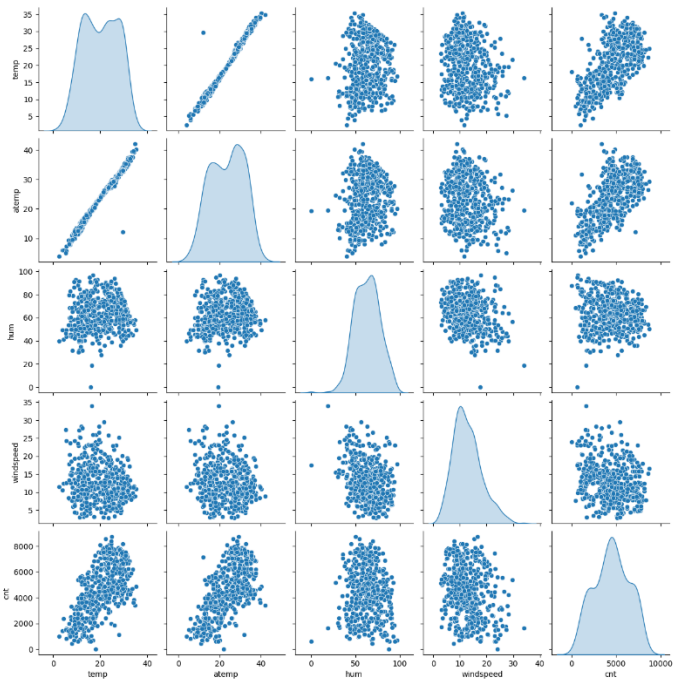
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
   Answer-
   Season, Month, Weekday , Weathersit, Holiday, Working day, Year are the categorical variables and the below conclusions are made with the help of box plot:-

- Season - Bike count is more in fall season(season 3), followed by season 2(summer) and season 4(winter) and is lowest in the spring season(season 1)
- Month – September has the highest count while December is the least. More bike count in the middle of the year and is a bit less in the beginning and end of the year.
- Weekday – Weekdays have high count than weekends.
- Weathersit - Most of the booking happened in weather1 (Clear, Few clouds, Partly cloudy, Partly cloudy) and the least in heavy rain/snow.
- Holiday - Almost 98% of the booking is made when it is not a holiday.
- Working day - Most of the booking happened on working day.
- Year -More number of bookings in 2019, and increase in business

2. Why is it important to use drop_first=True during dummy variable creation?

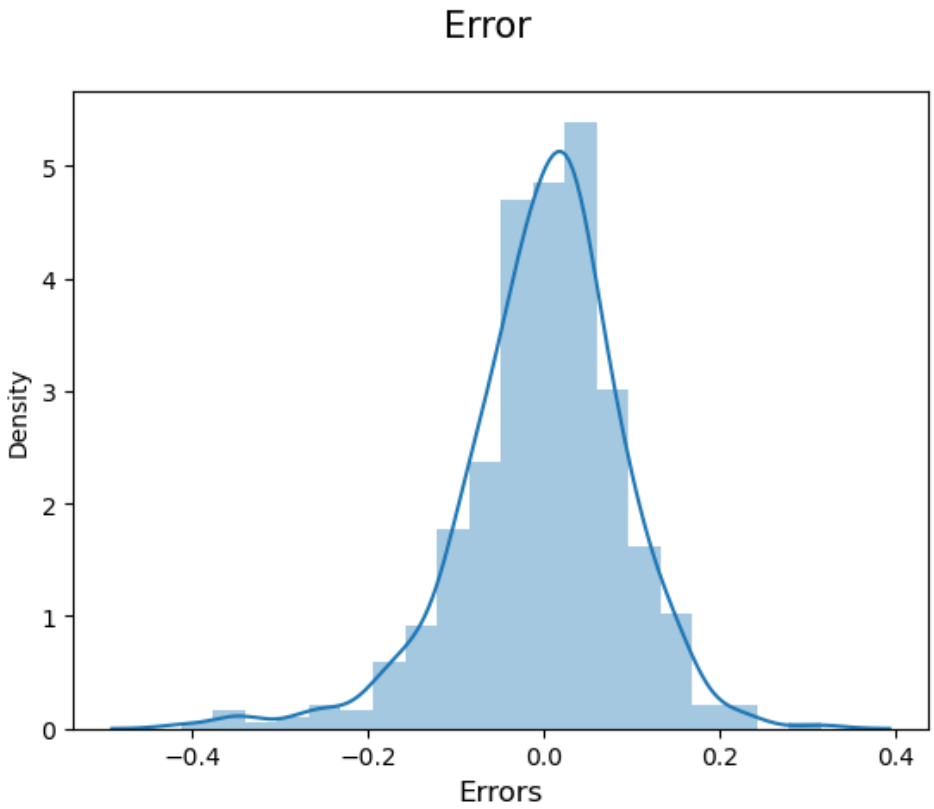
- If the drop_first= True is not used, then it leads to the creation of n dummies for n levels of categorical variables which results in increase in mutli collinearity between the variables. This also ensures that the model can distinguish between the effects of each category and prevents over-parameterization.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- 

Based on the above pairplot, temp and atemp are the numerical variables that are highly correlated with the target variable(cnt)

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- 

Based on the above distplot, it can be observed that the errors are normally distributed and the mean is centred over zero making the assumptions of Linear regression are valid.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- **Temperature (temp)** - A coefficient value of '0.5499' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5499 units.
- **Weather Situation 3 (weathersit_3)** - A coefficient value of -0.2880' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.2880 units.
- **Year (yr)** - A coefficient value of '0.2331' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2331 units.

General Subjective Questions

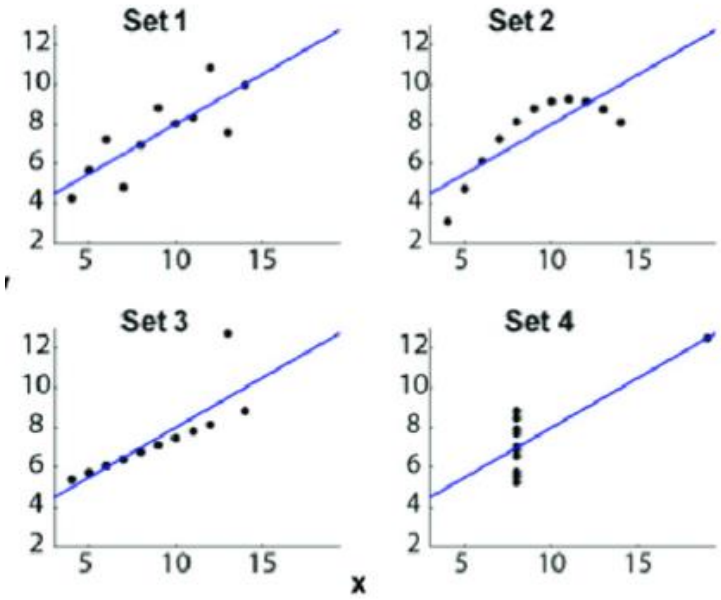
1. Explain the linear regression algorithm in detail.

- Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).
- Mathematically the equation for Linear regression is –
  y = mx+c

  Regression is divided into Simple Linear Regression and Multiple Linear Regression.
  Simple Linear Regression – when the target variable is dependent on only one variable.
  Multiple Linear Regression – when the target variable is dependent on more than one variable.

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet – It comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.



-

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).

- Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
- Data-set IV — looks like the value of x remains constant, except for one outlier as well.

3. What is Pearson's R?

- Pearson correlation coefficient, also known as Pearson R statistical test, measures the strength between the different variables and their relationships.
- Pearson's correlation coefficient can range from the value +1 to the value -1, where +1 indicates the perfect positive relationship between the variables considered, -1 indicates the perfect negative relationship between the variables considered, and 0 value indicates that no relationship exists between the variables considered.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Scaling - It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Why Scaling- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
- It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- VIF- A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis.
- The value of VIF is infinite means there is a perfect correlation.
- The formula for VIF is:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

- Q-Q plot – It is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against
- Use and importance - The purpose of the quantile-quantile (QQ) plot is to show if two data sets come from the same distribution. Plotting the first data set's quantiles along the x-axis

The Quantile-Quantile plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behaviour.